DOCUMENT RESUME

ED 210 310 TM 810 962

AUTHOR Stutzman, Thomas M.; And Others
TITLE Toward a Definition of "Good" Incident Dimension Congruence for BARS.
PUB DATE Mar 80
NOTE 21p.; Paper presented at the Annual Meeting of the Southeastern Psychological Association (Washington, DC, March, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Correlation; *Factor Structure; Higher Education; Questionnaires; *Rating Scales; Teacher Evaluation
IDENTIFIERS *Behaviorally Anchored Rating Scales; *Dimensional Analysis

ABSTRACT

This study investigated the relationship between the dimensionality of behaviorally anchored scales during construction and during their use. It was hypothesized that as the agreement among raters in assigning behaviors to dimensions increased, there would be an increase in the stability of the structure of the dimensions and in their independence during use. Undergraduate students enrolled in two introductory history classes and five introductory psychology classes were administered a teacher evaluation instrument containing 60 behaviorally based items. These items represented five performance dimensions and four ranges of rater agreement. For the analyses, class means were subtracted for each item, items were factor analyzed, and inter-dimension correlations were computed. The results demonstrated partial support for the hypothesis. At the higher levels of agreement among raters, more of the original five dimensions were interpretable, and they were more independent. However, in none of the factor analyses were all of the original dimensions reproduced. It was concluded that the degree of agreement among judges during dimension construction did influence the ability of factor analytic solutions to represent the original dimensionality. However, high agreement does not guarantee the maintenance of dimension independence. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED210310

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Toward a Definition of "Good" Incident Dimension Congruence for BARS

(A Comparison of Rational and Empirical Performance Dimensions)

Thomas M. Stutzman, William I. Sauser, Jr., and Samuel B. Green
Auburn University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
T.M. Stutzman
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the meeting of the Southeastern Psychological Association, Annual Washington, D.C., March 1980.

TM 810 962

Behaviorally Anchored Rating Scales (BARS) have received much attention since originally described by Smith and Kendall (1963). One reason for their popularity is that the BARS development process includes groups of future raters to (a) specify dimensions, (b) generate behaviors appropriate for each dimension, and (c) evaluate the behaviors through "retranslation". One judgment in the retranslation step is the reassignment of behaviors to dimensions. The P statistic represents the percentage of raters who assigned a behavior to the same dimension, i.e., the percentage of rater agreement. The higher the value of P, the more the raters view that behavior as appropriate for a single dimension.

Much of the attention received by BARS is due to the expectation that the retranslation process results in multiple, independent dimensions. This expectation has been questioned. Although it is obvious that multiple dimensions are generated, they are typically not independent in use. Schwab, Heneman and DeCotus (1975) reviewed a number of studies which investigated dimension independence and concluded that the evidence does not support the notion of multiple independent dimensions. Although the BARS' dimensions are sometimes less correlated than those associated with other measurement devices, the intercorrelations between BARS dimensions are substantial.

Schwab, et.al. suggested that the failure to obtain independent dimensions may be caused by the lax scaling criteria generally used in BARS construction. Specifically, these authors were referring to the low

percentage of agreement among judges in the placement of items on dimensions. In their classic article, Smith and Kendall (1963) suggested that incidents should be retained only if at least 60% of the judges agreed upon the assignment of an incident to a dimension.

In the one study where the P value was studied, Bernardin, LaShells, Smith and Alvares (1976) did not report the correlations between dimensions under either of their two P ranges (50-60 and 80 or greater). In their one comparison, they did not find any differences in rating variability, "the extent to which raters place a given ratee at the same level on different dimensions" (p. 76). Even so, they reaffirmed that the P criterion should be $\geq$ 60 percent.

Not only has dimension independence been questioned, but the meaning of the dimensions generated in the retranslation process has also become a topic. A number of studies have factor analyzed ratings using variations of BARS. Typically, the BARS are converted to mixed standard scales, where at least three ratings are obtained for each dimension. In no study to date have the original dimensions been reproduced.

Blanz and Ghiselli (1972) factor analyzed ratings on 18 retranslated dimensions with the result being four factors; Finely, Osburn, Dubin and Jeanneret (1977) factor analyzed 11 dimensions and obtained two and three factor solutions; and Kavanaugh and Duffy (1978) were required to interpret seven factors when they began with five dimensions.

Since the agreement statistic or P value can be viewed as an index of dimension overlap, and since cutoffs of 60 to 80 percent agreement are typically used in BARS construction, it is not surprising that overlap among dimensions is found or that factor structures differ from retranslated dimensions.

This issue is especially interesting since the development of BARS dimensions is a rational, judgmental process while the evaluation of dimension independence when BARS are used is an empirical process. The results of the two have differed to this point and it could be concluded that either one or the other procedure (or even both) are inaccurate and not valid. However, until a true test of their incongruence is made, any such conclusions would be premature. Such a test can only be performed when the dimensions are constructed to be independent.

The purpose of this study was to explore the independence of retranslated dimensions and to study the factor structure of retranslated dimensions when the P criterion associated with parallel sets of dimensions was varied. It was hypothesized that there would be less inter-dimension correlation and that factor analytic solutions would better represent the original BARS dimensions as the values of the P statistic were increased.

Method

A. Sample

Three-hundred sixty-three undergraduate students enrolled in two introductory history classes and five introductory psychology classes at Auburn served as participants. Students in the two history classes received no form of incentive for their participation.

B. Instrument

A teacher evaluation questionnaire containing 60 behaviorally based items was constructed. The items were selected from a large pool (Sauser, Evans, and Champion, 1979) which had been generated in a previous study using the retranslation technique (Sauser, 1979). Each of five dimensions were represented by 12 items: (a) relationship with students; (b) ability to present the material; (c) interest in course and material; (d) reasonableness of the workload; and (e) fairness of testing and grading. For each dimension, the items were selected so that four would have high scale values, four would have medium scale values and four would have low scale values. Further, one item for each dimension representing each scale value was chosen so that four ranges of P values (60-69, 70-79, 80-89 and 90-100) were equally represented.

The instructions for the questionnaire specified that the raters were to evaluate the teacher of the class they were in, using five available options. The resulting format of the rating scale is a slight modification of the mixed-standard rating scale first described

by Blanz and Ghiselli (1972). In this procedure, raters compare a behavioral statement with the behavior of the ratee and determine if the ratee performs at a level better than the example, at the same level as the example, or at a level below the example. For this study, more specificity was attempted. Instead of the three traditional response options, five response options were used: much better, a little better, about the same, a little worse, and much worse.

C. Statistical Analyses

Initially the items were scored from "1" to "5" with five representing the highest performance rating. Before the dimension scores were completed, the five point scale was converted to a three-point scale and each response was weighted so as to conform with the mixed-standard scoring procedure suggested by Saal (1979). Using this procedure, the original "1" and "2" responses were grouped together and scored "1", the "3" response was scored "2", and the "4" and "5" responses were grouped together and scored "3". The new three-point scale responses were then differentially weighted as follows: for the high scale items, the "1", "2", and "3" responses were weighted 7, 8 and 9 respectively; the medium scale items were weighted 4, 5, 6; and the low scale items were weighted 1, 2, 3, i.e., they were unchanged. Each set of three items was then summed and eleven was subtracted from the sum to produce 20 dimension means, five for each P range. These dimension means fell along a seven point continuum.

Prior to the computation of correlations among the dimensions, or the factor analyses of the items, class means were subtracted

from all item and dimension scores to eliminate differences across classes. To test for dimension independence, Pearson-product moment correlations were computed between the dimensions, the correlations were transferred to Fisher's Z scores, and mean dimension correlations were computed at each P value.

In an effort to reproduce the original dimensionality of the data, within groups factor analysis (Linn, Centra and Tucker, 1975) was performed with each set of 15 items. The varimax orthogonal procedure with squared multiple correlations in the diagonal, was followed by an Oblique Promax rotation. Initially, five factors were extracted since five dimensions created through retranslation were represented.

Results

Dimension Independence And Dimension Use

To determine the extent to which the dimensions were used independently, two analyses were performed. The first analysis consisted of comparing the mean dimension scores within each P value. The dimension means are included in Table 1, and as can be seen, the mean dimension scores did vary as a function of the dimension rated. Unfortunately, they varied even more as a function of the P value.

The second analysis consisted of correlating each dimension with the other four dimensions and then computing the average within P value, between dimension correlation. Table 2 contained the results of this analysis. While lower between dimension correlations were expected for the higher P values, the expected result occurred only

only for the highest P level (90's). The finding of a reasonably low between dimensions average correlation with the lowest P value (60) questions the hypothesized relationship between P values and dimension independence.

Tables 1 & 2 Here

Rational and Empirical Dimension Structure

A summary of the results of all of the factor analyses is presented in Table 3. As can be seen, only when the P values were 90 or greater were five factors indicated by the Kaiser Criterion (eigen values > 1). Similarly, only when the P values were 90 or greater were three or more content factors interpretable. Unfortunately the resulting factor structures were heavily influenced by the scale values as well as the content dimension. Thus some of the content dimensions were not reproduced in the factor analysis because some of the factors (usually two) are strong scale level factors.

The results of a sample of the factor analyses are included in Tables 4, 5, 6, 7 and 8. In Table 4, three of the original dimensions are reproduced (A, B and E), although the two factors which account for the majority of the variance are scale factors. As the P values decrease from 90, the number of interpretable content factors also decrease with no content factors found for the lowest P values (Table 7).

In an attempt to separate the scale factors from the content factors, seven and eight factor solutions were tried. As can be seen

in Table 8, this attempt met with little success. The content factors continued to dominate the results and the number of content factors never exceeded the three found with the P values of 90 or greater.

Tables 4-8 Here

Discussion

Both the results of the factor analyses and the correlational analyses appear to support the original hypothesis that P values influence the independence of BARS dimensions when they are used. Unfortunately the extent of this influence was not quantifiable due to the presence of scale factors and the lack of an available confirmator factor analysis procedure.

The weak evidence presented should serve as an illustration of the caution given by Kavanaugh and Duffy, "The psychometric evidence derived during the development of the retranslated dimensions is a necessary, but not sufficient condition for the measurement of the scaling dimensions" (1978, p. 470). High P values do not guarantee dimension independence nor do they guarantee simple structure.

The incongruence between the independence of dimensions during their construction and their use should serve as a catalyst for future work. It is quite discouraging to denote extensive efforts in the development of precise measuring instruments only to discover that they are not working or at least to discover that they are not being used as

expected. Although one can always blame the problem on the raters, it is still puzzling that the raters can develop a set of independent dimensions but they are unable to use them independently.

References

Bernardin, H. J., LaShells, M. B., Smith, P. C. and Alvares, K. M. Behavioral expectations scales: Effects of developmental procedures and formats. Journal of Applied Psychology, 1976, 61, 75-79.

Blanz, F. and Ghiselli, E. E. The mixed standard scale: A new rating system. Personnel Psychology, 1972, 25, 185-199.

Finley, D. M., Osburn, H. G., Dubin, J. A. and Jeanneret, P. R. Behaviorally based rating scales: Effects of specific anchors and disguised scale continua. Personnel Psychology, 1977, 30, 659-669.

Kavanaugh, M. J. and Duffy, J. F. An extension and field test of the retranslation method for developing rating scales. Personnel Psychology, 1978, 31, 461-470.

Linn, R. L., Centra, J. A. and Tucker, L. Between, within, and total group factor analyses of student ratings of instruction. Multivariate Behavioral Research, 1975, 10, 227-288.

Saal, F. E. Mixed standard rating scale: A consistent system for numerically coding inconsistent response combinations. Journal of Applied Psychology, 1979, 64, 422-428.

Sauser, W. I., Jr. A comparative evaluation of the effects of rater participation and rater training on characteristics of employee performance appraisal ratings and related mediating variables (Doctoral dissertation, Georgia Institute of Technology, 1978). Dissertation Abstracts International, 1979, 39, 5116-B.

Sauser, W. I., Jr., Evans, K. L. and Champion, L. C. H. Two-hundred and fifty scaled incidents of college classroom teaching behavior. Paper presented at the meeting of the Southeastern Psychological Association, New Orleans, March 1979.

Schwab, D. P., Heneman, H. G., III. and DeCotus, T. A. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 1975, 28, 549-562.

Smith, P. C. and Kendall, L. M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Table 1

Dimension Means X P Value (N = 321)

Dimension	P Value			
	90's	80's	70's	60's
A	5.44	4.41	5.13	5.32
B	4.42	5.49	5.62	4.97
C	5.61	4.81	5.30	5.33
D	5.87	4.81	4.81	5.66
E	4.39	5.87	5.07	4.91

Table 2

Average Between Dimension Correlations by P Value[a]

Dimension	90's	80's	70's	60's
A	.33	.49	.37	.46
B	.34	.50	.44	.44
C	.36	.46	.45	.40
D	.34	.36	.38	.37
E	.32	.44	.41	.33
Average Inter-Dimension Correlations	.34	.45	.41	.38

[a]All coefficients were transformed using Fisher's r to z transformation prior to the calculation of the means.

Table 3

Summary of Promax Results

	No. of Content Factors	% Var Accounted For	Kaiser Criterion (No. of Eigen Values ≥ 1)
5 Factor Solution			
90's	3	62%	5
80's	1	62%	4
70's	1	58%	4
60's	0	58%	4
7 Factor Solution			
90's	3	73%	5
80's	2	72%	4
70's	1	69%	4
60's	1	69%	4
8 Factor Solution			
90's	3	79%	5
80's	2	77%	4
70's	0	74%	4
60's	0	75%	4

Table 4

Promax Factor Structure Matrix
Variables With P Values ≥ 90, Loadings ≥ |40|
The Five Factors Account for 62% of the Total Variance

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
A93		-.56			
92					.61
91					.63
B93		-.62			
92		-.61		.43	
91	-.46			.92	
C93		-.54			
92	-.58				.47
91	-.61				.46
D93		-.49			
92	-.46				
91	-.61			.40	
E93			-.84		
92			-.61		
91	-.70				
Factor Interpretation	Low Scale Values	High Scale Values	Dimension E	Dimension B	Dimension A

Table 5

Promax Factor Structure Matrix
Variables With P Values ≥ 80 and < 90, Loadings ≥ |40|
The Five Factors Accounted for 62% of the Total Variance

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
A83		-54	41		
82		-55			
81	-67	-45			
B83		-47			
82				57	
81	-61				
C83			81		
82			57		
81	-81			41	
D83		-53			
82					
81	-62				
E83		-60			
82	-68				
81	-51			55	
Factor Interpretation	Low Scale Factor	High Scale Factor	Dimension C	Low Scale Factor	

Table 6

Promax Factor Structure Matrix
Variables With P Values ≥ 70 and < 80, Loadings ≥ |40|
The Five Factors Accounted for 58% of the Total Variance

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
A73		-52			
72					
71	-73			58	
B73		-53			
72				56	
71	-56			48	
C73		-42			48
72	-49			46	47
71	-45			77	
D73		-60			
72					
71	-70			48	
E73		-50			
72			-64		
71				61	
Factor Interpretation	Low Scale Factor	High Scale Factor		Low Scale Factor	Dimension C

Table 7

Promax Factor Structure Matrix
Variables With P Values ≥ 60 and < 70, Loadings ≥ |40|
The Five Factors Accounted for 58% of the Total Variance

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
A63		-58			
62	53		-43		
61	86				
B63		-56			
62			-61		
61	50		-63		
C63		-54			
62					
61			-70		
D63					
62			-41		
61	65				
E63					
62				55	
61	51		-45		
Factor Interpretation	Low Scale Factor	High Scale Factor	Low Scale Factor		

Table 8

Promax Factor Structure Matrix
Variables With P Values ≥ 90, Loadings ≥ |40|
The Seven Factors Account For 73% Of The Total Variance

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
A93		-58					
92	-44					74	
91						59	
B93		-61					
92		-59		47			
91				88			
C93		-53					
92	-71					40	
91	-69						
D93		-51					
92					67		
91	-46						-69
E93		-40	-76				
92			-72				
91	-45				45		-66
Factor Interpretation	Low Scale Factor	High Scale Factor	Dimension E	Dimension B	Low Scale Factor	Dimension A	Low Scale Factor